SAS/STAT® User's Guide
The CANDISC Procedure

2022.12*

* This document might apply to additional versions of the software. Open this document in SAS Help Center and click on the version in the banner to see all available versions.

SAS® Documentation
December 15, 2022

This document is an individual chapter from *SAS/STAT® User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2022. *SAS/STAT® User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® User's Guide

December 2022

Chapter 35
The CANDISC Procedure

Contents

Overview: CANDISC Procedure

Canonical discriminant analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. The methodology that is used in deriving the canonical coefficients parallels that of a one-way multivariate analysis of variance (MANOVA). MANOVA tests for equality of the mean vector across class levels. Canonical discriminant analysis finds linear combinations of the quantitative variables that provide maximal separation between classes or groups. Given a classification variable and several quantitative variables, the CANDISC procedure derives *canonical variables*, which are linear combinations of the quantitative variables that summarize between-class variation in much the same way that principal components summarize total variation.

The CANDISC procedure performs a canonical discriminant analysis, computes squared Mahalanobis distances between class means, and performs both univariate and multivariate one-way analyses of variance. Two output data sets can be produced: one that contains the canonical coefficients and another that contains, among other things, scored canonical variables. You can rotate the canonical coefficients by using the FACTOR procedure. It is customary to standardize the canonical coefficients so that the canonical variables have means that are equal to 0 and pooled within-class variances that are equal to 1. PROC CANDISC displays both standardized and unstandardized canonical coefficients. Correlations between the canonical variables and the original variables in addition to the class means for the canonical variables are also displayed; these correlations, sometimes known as loadings, are called canonical structures. To aid the visual interpretation of group differences, you can use ODS Graphics to display graphs of pairs of canonical variables from the scored canonical variables output data set.

When you have two or more groups of observations that have measurements on several quantitative variables, canonical discriminant analysis derives a linear combination of the variables that has the highest possible multiple correlation with the groups. This maximal multiple correlation is called the *first canonical correlation*. The coefficients of the linear combination are the *canonical coefficients* or *canonical weights*. The variable that is defined by the linear combination is the *first canonical variable* or *canonical component*. The second canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlation with the groups. The process of extracting canonical variables can be repeated until the number of canonical variables equals the number of original variables or the number of classes minus one, whichever is smaller.

The first canonical correlation is at least as large as the multiple correlation between the groups and any of the original variables. If the original variables have high within-group correlations, the first canonical correlation can be large even if all the multiple correlations are small. In other words, the first canonical variable can show substantial differences between the classes, even if none of the original variables do. Canonical variables are sometimes called *discriminant functions*, but this usage is ambiguous because the DISCRIM procedure produces very different functions for classification that are also called discriminant functions.

For each canonical correlation, PROC CANDISC tests the hypothesis that it and all smaller canonical correlations are zero in the population. An F approximation (Rao 1973; Kshirsagar 1972) is used that gives better small-sample results than the usual chi-square approximation. The variables should have an approximate multivariate normal distribution within each class, with a common covariance matrix in order for the probability levels to be valid.

Canonical discriminant analysis is equivalent to canonical correlation analysis between the quantitative variables and a set of dummy variables coded from the CLASS variable. Performing canonical discriminant analysis is also equivalent to performing the following steps:

1. Transform the variables so that the pooled within-class covariance matrix is an identity matrix.
2. Compute class means on the transformed variables.
3. Perform a principal component analysis on the means, weighting each mean by the number of observations in the class. The eigenvalues are equal to the ratio of between-class variation to within-class variation in the direction of each principal component.
4. Back-transform the principal components into the space of the original variables to obtain the canonical variables.

An interesting property of the canonical variables is that they are uncorrelated whether the correlation is calculated from the total sample or from the pooled within-class correlations. However, the canonical coefficients are not orthogonal, so the canonical variables do not represent perpendicular directions through the space of the original variables.

Getting Started: CANDISC Procedure

The data in this example are measurements of 159 fish caught in Finland's Lake Laengelmaevesi; this data set is available from the Puranen. For each of the seven species (bream, roach, whitefish, parkki, perch, pike, and smelt), the weight, length, height, and width of each fish are tallied. Three different length measurements are recorded: from the nose of the fish to the beginning of its tail, from the nose to the notch of its tail, and from the nose to the end of its tail. The height and width are recorded as percentages of the third length variable. The fish data set is available from the Sashelp library.

The following step uses PROC CANDISC to find the three canonical variables that best separate the species of fish in the Sashelp.Fish data and create the output data set outcan. When the NCAN=3 option is specified, only the first three canonical variables are displayed. The ODS EXCLUDE statement excludes the canonical structure tables and most of the canonical coefficient tables in order to obtain a more compact set of results. The TEMPLATE and SGRENDER procedures create a plot of the first two canonical variables. The following statements produce Figure 35.1 through Figure 35.6:

```
title 'Fish Measurement Data';

proc candisc data=sashelp.fish ncan=3 out=outcan;
   ods exclude tstruc bstruc pstruc tcoef pcoef;
   class Species;
   var Weight Length1 Length2 Length3 Height Width;
run;

proc template;
   define statgraph scatter;
      begingraph / attrpriority=none;
         entrytitle 'Fish Measurement Data';
         layout overlayequated / equatetype=fit
            xaxisopts=(label='Canonical Variable 1')
            yaxisopts=(label='Canonical Variable 2');
            scatterplot x=Can1 y=Can2 / group=species name='fish'
```

```
                                          markerattrs=(size=3px);
            layout gridded / autoalign=(topright);
               discretelegend 'fish' / border=false opaque=false;
            endlayout;
         endlayout;
      endgraph;
   end;
run;

proc sgrender data=outcan template=scatter;
run;
```

PROC CANDISC begins by displaying summary information about the variables in the analysis. This information includes the number of observations, the number of quantitative variables in the analysis (specified with the VAR statement), and the number of classes in the classification variable (specified with the CLASS statement). The frequency of each class is also displayed.

Figure 35.1 Summary Information

Fish Measurement Data

The CANDISC Procedure

Total Sample Size	158	**DF Total**	157
Variables	6	**DF Within Classes**	151
Classes	7	**DF Between Classes**	6

Number of Observations Read	159
Number of Observations Used	158

Class Level Information

Species	Variable Name	Frequency	Weight	Proportion
Bream	Bream	34	34.0000	0.215190
Parkki	Parkki	11	11.0000	0.069620
Perch	Perch	56	56.0000	0.354430
Pike	Pike	17	17.0000	0.107595
Roach	Roach	20	20.0000	0.126582
Smelt	Smelt	14	14.0000	0.088608
Whitefish	Whitefish	6	6.0000	0.037975

PROC CANDISC performs a multivariate one-way analysis of variance (one-way MANOVA) and provides four multivariate tests of the hypothesis that the class mean vectors are equal. These tests, shown in Figure 35.2, indicate that not all the mean vectors are equal ($p < 0.0001$).

Figure 35.2 MANOVA and Multivariate Tests

Fish Measurement Data

The CANDISC Procedure

Multivariate Statistics and F Approximations

S=6 M=-0.5 N=72

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00036325	90.71	36	643.89	<.0001
Pillai's Trace	3.10465132	26.99	36	906	<.0001
Hotelling-Lawley Trace	52.05799676	209.24	36	413.64	<.0001
Roy's Greatest Root	39.13499776	984.90	6	151	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

The first canonical correlation is the greatest possible multiple correlation with the classes that can be achieved by using a linear combination of the quantitative variables. The first canonical correlation, displayed in Figure 35.3, is 0.987463. Figure 35.3 also displays a likelihood ratio test of the hypothesis that the current canonical correlation and all smaller ones are zero. The first line is equivalent to Wilks' lambda multivariate test.

Figure 35.3 Canonical Correlations

Fish Measurement Data

The CANDISC Procedure

					Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalue	Difference	Proportion	Cumulative
1	0.987463	0.986671	0.001989	0.975084	39.1350	29.3859	0.7518	0.7518
2	0.952349	0.950095	0.007425	0.906969	9.7491	7.3786	0.1873	0.9390
3	0.838637	0.832518	0.023678	0.703313	2.3706	1.7016	0.0455	0.9846
4	0.633094	0.623649	0.047821	0.400809	0.6689	0.5346	0.0128	0.9974
5	0.344157	0.334170	0.070356	0.118444	0.1344	0.1343	0.0026	1.0000
6	0.005701	.	0.079806	0.000033	0.0000		0.0000	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.00036325	90.71	36	643.89	<.0001
2	0.01457896	46.46	25	547.58	<.0001
3	0.15671134	23.61	16	452.79	<.0001
4	0.52820347	12.09	9	362.78	<.0001
5	0.88152702	4.88	4	300	0.0008
6	0.99996749	0.00	1	151	0.9442

The first canonical variable, Can1, shows that the linear combination of the centered variables Can1 = $-0.0006 \times$ Weight $- 0.33 \times$ Length1 $2.49 \times$ Length2 $+ 2.60 \times$ Length3 $+ 1.12 \times$ Height $- 1.45 \times$ Width separates the species most effectively (see Figure 35.4).

Figure 35.4 Raw Canonical Coefficients

Fish Measurement Data

The CANDISC Procedure

Raw Canonical Coefficients			
Variable	**Can1**	**Can2**	**Can3**
Weight	-0.000648508	-0.005231659	-0.005596192
Length1	-0.329435762	-0.626598051	-2.934324102
Length2	-2.486133674	-0.690253987	4.045038893
Length3	2.595648437	1.803175454	-1.139264914
Height	1.121983854	-0.714749340	0.283202557
Width	-1.446386704	-0.907025481	0.741486686

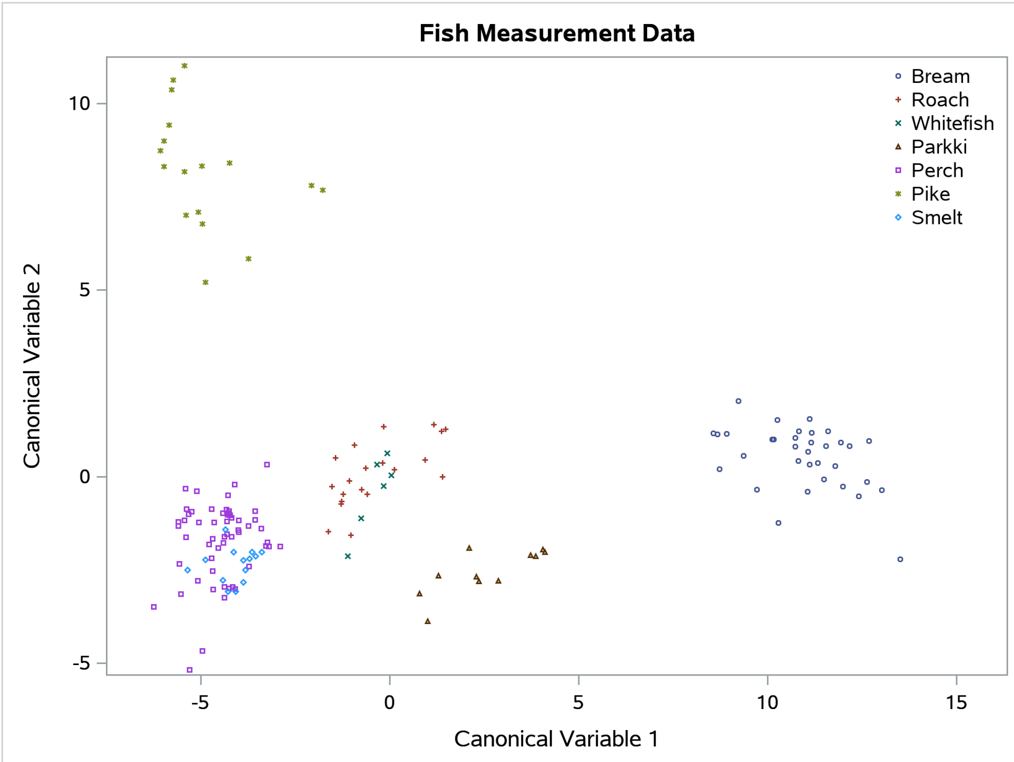
PROC CANDISC computes the means of the canonical variables for each class. The first canonical variable is the linear combination of the variables Weight, Length1, Length2, Length3, Height, and Width that provides the greatest difference (in terms of a univariate F test) between the class means. The second canonical variable provides the greatest difference between class means while being uncorrelated with the first canonical variable.

Figure 35.5 Class Means for Canonical Variables

Class Means on Canonical Variables			
Species	**Can1**	**Can2**	**Can3**
Bream	10.94142464	0.52078394	0.23496708
Parkki	2.58903743	-2.54722416	-0.49326158
Perch	-4.47181389	-1.70822715	1.29281314
Pike	-4.89689441	8.22140791	-0.16469132
Roach	-0.35837149	0.08733611	-1.10056438
Smelt	-4.09136653	-2.35805841	-4.03836098
Whitefish	-0.39541755	-0.42071778	1.06459242

Figure 35.6 displays a plot of the first two canonical variables, which shows that Can1 discriminates among three groups: (1) bream; (2) whitefish, roach, and parkki; and (3) smelt, pike, and perch. Can2 best discriminates between pike and the other species.

Figure 35.6 Plot of First Two Canonical Variables



Syntax: CANDISC Procedure

The following statements are available in the CANDISC procedure:

> **PROC CANDISC** < *options* > ;
> 	**CLASS** *variable* ;
> 	**BY** *variables* ;
> 	**FREQ** *variable* ;
> 	**VAR** *variables* ;
> 	**WEIGHT** *variable* ;

The BY, CLASS, FREQ, VAR, and WEIGHT statements are described in alphabetical order after the PROC CANDISC statement.

PROC CANDISC Statement

PROC CANDISC < *options* > **;**

The PROC CANDISC statement invokes the CANDISC procedure. Table 35.1 summarizes the options available in the PROC CANDISC statement.

Table 35.1 CANDISC Procedure Options

Option	Description
Input Data Set	
DATA=	Specifies the input SAS data set
Output Data Sets	
OUT=	Specifies the output data set that contains the canonical scores
OUTSTAT=	Specifies the output statistics data set
Method Details	
NCAN=	Specifies the number of canonical variables
PREFIX=	Specifies a prefix for naming the canonical variables
SINGULAR=	Specifies the singularity criterion
Control Displayed Output	
ALL	Displays all output
ANOVA	Displays univariate statistics
BCORR	Displays between correlations
BCOV	Displays between covariances
BSSCP	Displays between SSCPs
DISTANCE	Displays squared Mahalanobis distances
NOPRINT	Suppresses all displayed output
PCORR	Displays pooled correlations
PCOV	Displays pooled covariances
PSSCP	Displays pooled SSCPs
SHORT	Suppresses some displayed output
SIMPLE	Displays simple descriptive statistics
STDMEAN	Displays standardized class means
TCORR	Displays total correlations
TCOV	Displays total covariances
TSSCP	Displays total SSCPs
WCORR	Displays within correlations
WCOV	Displays within covariances
WSSCP	Displays within SSCPs

ALL
activates all the display options.

ANOVA
displays univariate statistics for testing the hypothesis that the class means are equal in the population for each variable.

BCORR
displays between-class correlations.

BCOV
displays between-class covariances. The between-class covariance matrix equals the between-class SSCP matrix divided by $n(c-1)/c$, where n is the number of observations and c is the number of classes. The between-class covariances should be interpreted in comparison with the total-sample and within-class covariances, not as formal estimates of population parameters.

BSSCP
displays the between-class SSCP matrix.

DATA=*SAS-data-set*
specifies the data set to be analyzed. The data set can be an ordinary SAS data set or one of several specially structured data sets created by SAS statistical procedures. These specially structured data sets include TYPE=CORR, TYPE=COV, TYPE=CSSCP, and TYPE=SSCP. If you omit the DATA= option, PROC CANDISC uses the most recently created SAS data set.

DISTANCE
MAHALANOBIS
displays squared Mahalanobis distances between the group means, the F statistics, and the corresponding probabilities of greater squared Mahalanobis distances between the group means.

NCAN=*n*
specifies the number of canonical variables to be computed. The value of *n* must be less than or equal to the number of variables. If you specify NCAN=0, the procedure displays the canonical correlations but not the canonical coefficients, structures, or means. A negative value suppresses the canonical analysis entirely. Let v be the number of variables in the VAR statement, and let c be the number of classes. If you omit the NCAN= option, only $\min(v, c-1)$ canonical variables are generated; if you also specify an OUT= output data set, v canonical variables are generated, and the last $v-(c-1)$ canonical variables have missing values.

NOPRINT
suppresses the normal display of results. This option temporarily disables the Output Delivery System (ODS). For more information about ODS, see Chapter 23, "Using the Output Delivery System."

OUT=*SAS-data-set*
creates an output SAS data set to contain the original data and the canonical variable scores. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see *SAS Programmers Guide: Essentials*.

OUTSTAT=*SAS-data-set*
creates a TYPE=CORR output SAS data set to contain various statistics, including class means, standard deviations, correlations, canonical correlations, canonical structures, canonical coefficients, and means of canonical variables for each class. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see *SAS Programmers Guide: Essentials*.

PCORR
displays pooled within-class correlations (partial correlations based on the pooled within-class covariances).

PCOV
displays pooled within-class covariances.

PREFIX=*name*
specifies a prefix for naming the canonical variables. By default the names are Can1, Can2, Can3, and so on. If you specify PREFIX=Abc, the components are named Abc1, Abc2, and so on. The number of characters in the prefix plus the number of digits required to designate the canonical variables should not exceed the length defined by the VALIDVARNAME= system option (for example, 32 for VALIDVARNAME=V7). The prefix is truncated if the combined length exceeds the maximum.

PSSCP
displays the pooled within-class corrected SSCP matrix.

SHORT
suppresses the display of canonical structures, canonical coefficients, and class means on canonical variables; only tables of canonical correlations and multivariate test statistics are displayed.

SIMPLE
displays simple descriptive statistics for the total sample and within each class.

SINGULAR=*p*
specifies the criterion for determining the singularity of the total-sample correlation matrix and the pooled within-class covariance matrix, where $0 < p < 1$. The default is SINGULAR=1E–8.

Let **S** be the total-sample correlation matrix. If the R square for predicting a quantitative variable in the VAR statement from the variables that precede it exceeds $1 - p$, then **S** is considered singular. If **S** is singular, the probability levels for the multivariate test statistics and canonical correlations are adjusted for the number of variables whose R square exceeds $1 - p$.

If **S** is considered singular and the inverse of **S** (squared Mahalanobis distances) is required, a quasi inverse is used instead. For more information, see the section "Quasi-inverse" on page 2735 in Chapter 42, "The DISCRIM Procedure."

STDMEAN
displays total-sample and pooled within-class standardized class means.

TCORR
: displays total-sample correlations.

TCOV
: displays total-sample covariances.

TSSCP
: displays the total-sample corrected SSCP matrix.

WCORR
: displays within-class correlations for each class level.

WCOV
: displays within-class covariances for each class level.

WSSCP
: displays the within-class corrected SSCP matrix for each class level.

BY Statement

BY *variables* **;**

You can specify a BY statement in PROC CANDISC to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement in the CANDISC procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the “Grouping Data” section of *SAS Programmers Guide: Essentials*. For more information about the DATASETS procedure, see the discussion in the *Base SAS Procedures Guide*.

CLASS Statement

CLASS *variable* ;

The values of the CLASS variable define the groups for analysis. Class levels are determined by the formatted values of the CLASS variable. The CLASS variable can be numeric or character. A CLASS statement is required.

FREQ Statement

FREQ *variable* ;

If a variable in the data set represents the frequency of occurrence of the other values in the observation, include the name of the variable in a FREQ statement. The procedure then treats the data set as if each observation appears n times, where n is the value of the FREQ variable for the observation. The total number of observations is considered to be equal to the sum of the FREQ variable when the procedure determines degrees of freedom for significance probabilities.

If the value of the FREQ variable is missing or is less than 1, the observation is not used in the analysis. If the value is not an integer, the value is truncated to an integer.

VAR Statement

VAR *variables* ;

You specify the quantitative variables to include in the analysis by using a VAR statement. If you do not use a VAR statement, the analysis includes all numeric variables not listed in other statements.

WEIGHT Statement

WEIGHT *variable* ;

To use relative weights for each observation in the input data set, place the weights in a variable in the data set and specify the name in a WEIGHT statement. This is often done when the variance associated with each observation is different and the values of the WEIGHT variable are proportional to the reciprocals of the variances. If the value of the WEIGHT variable is missing or is less than 0, then a value of 0 for the weight is assumed.

The WEIGHT and FREQ statements have a similar effect except that the WEIGHT statement does not alter the degrees of freedom.

Details: CANDISC Procedure

Missing Values

If an observation has a missing value for any of the quantitative variables, it is omitted from the analysis. If an observation has a missing CLASS value but is otherwise complete, it is not used in computing the canonical correlations and coefficients; however, canonical variable scores are computed for that observation for the OUT= data set.

Computational Details

General Formulas

Canonical discriminant analysis is equivalent to canonical correlation analysis between the quantitative variables and a set of dummy variables coded from the CLASS variable. In the following notation, the dummy variables are denoted by $\mathbf{y}$ and the quantitative variables are denoted by $\mathbf{x}$. The total sample covariance matrix for the $\mathbf{x}$ and $\mathbf{y}$ variables is

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{bmatrix}$$

When c is the number of groups, n_t is the number of observations in group t, and $\mathbf{S}_t$ is the sample covariance matrix for the $\mathbf{x}$ variables in group t, the within-class pooled covariance matrix for the $\mathbf{x}$ variables is

$$\mathbf{S}_p = \frac{1}{\sum n_t - c} \sum (n_t - 1)\mathbf{S}_t$$

The canonical correlations, ρ_i, are the square roots of the eigenvalues, λ_i, of the following matrix. The corresponding eigenvectors are $\mathbf{v}_i$.

$$\mathbf{S}_p{}^{-1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}{}^{-1}\mathbf{S}_{yx}\mathbf{S}_p{}^{-1/2}$$

Let $\mathbf{V}$ be the matrix that contains the eigenvectors $\mathbf{v}_i$ that correspond to nonzero eigenvalues as columns. The raw canonical coefficients are calculated as follows:

$$\mathbf{R} = \mathbf{S}_p{}^{-1/2}\mathbf{V}$$

The pooled within-class standardized canonical coefficients are

$$\mathbf{P} = \mathrm{diag}(\mathbf{S}_p)^{1/2}\mathbf{R}$$

The total sample standardized canonical coefficients are

$$\mathbf{T} = \mathrm{diag}(\mathbf{S}_{xx})^{1/2}\mathbf{R}$$

Let $\mathbf{X}_c$ be the matrix that contains the centered $\mathbf{x}$ variables as columns. The canonical scores can be calculated by any of the following:

$$\mathbf{X}_c\,\mathbf{R}$$

$$\mathbf{X}_c \operatorname{diag}(\mathbf{S}_p)^{-1/2}\mathbf{P}$$

$$\mathbf{X}_c \operatorname{diag}(\mathbf{S}_{xx})^{-1/2}\mathbf{T}$$

For the multivariate tests based on $\mathbf{E}^{-1}\mathbf{H}$,

$$\mathbf{E} = (n-1)(\mathbf{S}_{yy} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy})$$

$$\mathbf{H} = (n-1)\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}$$

where n is the total number of observations.

Input Data Set

The input DATA= data set can be an ordinary SAS data set or one of several specially structured data sets created by statistical procedures available in SAS/STAT software. For more information about special types of data sets, see Appendix A, "Special SAS Data Sets." The BY variable in these data sets becomes the CLASS variable in PROC CANDISC. These specially structured data sets include the following:

- TYPE=CORR data sets created by PROC CORR by using a BY statement
- TYPE=COV data sets created by PROC PRINCOMP by using both the COV option and a BY statement
- TYPE=CSSCP data sets created by PROC CORR by using the CSSCP option and a BY statement, where the OUT= data set is assigned TYPE=CSSCP by using the TYPE= data set option
- TYPE=SSCP data sets created by PROC REG by using both the OUTSSCP= option and a BY statement

When the input data set is TYPE=CORR, TYPE=COV, or TYPE=CSSCP, then PROC CANDISC reads the number of observations for each class from the observations for which _TYPE_='N' and the variable means in each class from the observations for which _TYPE_='MEAN'. The CANDISC procedure then reads the within-class correlations from the observations for which _TYPE_='CORR', the standard deviations from the observations for which _TYPE_='STD' (data set TYPE=CORR), the within-class covariances from the observations for which _TYPE_='COV' (data set TYPE=COV), or the within-class corrected sums of squares and crossproducts from the observations for which _TYPE_='CSSCP' (data set TYPE=CSSCP).

When the data set does not include any observations for which _TYPE_='CORR' (data set TYPE=CORR), _TYPE_='COV' (data set TYPE=COV), or _TYPE_='CSSCP' (data set TYPE=CSSCP) for each class, PROC CANDISC reads the pooled within-class information from the data set. In this case, PROC CANDISC reads the pooled within-class correlations from the observations for which _TYPE_='PCORR', the pooled within-class standard deviations from the observations for which _TYPE_='PSTD' (data set TYPE=CORR), the pooled within-class covariances from the observations for which _TYPE_='PCOV' (data set TYPE=COV), or the pooled within-class corrected SSCP matrix from the observations for which _TYPE_='PSSCP' (data set TYPE=CSSCP).

When the input data set is TYPE=SSCP, then PROC CANDISC reads the number of observations for each class from the observations for which _TYPE_='N', the sum of weights of observations from the variable Intercept in observations for which _TYPE_='SSCP' and _NAME_='Intercept', the variable sums from the analysis variables in observations for which _TYPE_='SSCP' and _NAME_='Intercept', and the uncorrected sums of squares and crossproducts from the analysis variables in observations for which _TYPE_='SSCP' and _NAME_=*variable-name*.

Output Data Sets

OUT= Data Set

The OUT= data set contains all the variables in the original data set plus new variables that contain the canonical variable scores. You determine the number of new variables by using the NCAN= option. The names of the new variables are formed as they are for the PREFIX= option. The new variables have means equal to 0 and pooled within-class variances equal to 1. An OUT= data set cannot be created if the DATA= data set is not an ordinary SAS data set.

OUTSTAT= Data Set

The OUTSTAT= data set is similar to the TYPE=CORR data set that the CORR procedure produces but contains many results in addition to those produced by the CORR procedure.

The OUTSTAT= data set is TYPE=CORR, and it contains the following variables:

- the BY variables, if any
- the CLASS variable
- _TYPE_, a character variable of length 8 that identifies the type of statistic
- _NAME_, a character variable that identifies the row of the matrix or the name of the canonical variable
- the quantitative variables (those in the VAR statement, or if there is no VAR statement, all numeric variables not listed in any other statement)

The observations, as identified by the variable _TYPE_, have the following _TYPE_ values:

TYPE	**Contents**
N	number of observations both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
SUMWGT	sum of weights both for the total sample (CLASS variable missing) and within each class (CLASS variable present) if a WEIGHT statement is specified
MEAN	means both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
STDMEAN	total-standardized class means
PSTDMEAN	pooled within-class standardized class means
STD	standard deviations both for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PSTD	pooled within-class standard deviations
BSTD	between-class standard deviations
RSQUARED	univariate R squares

The following kinds of observations are identified by the combination of the variables _TYPE_ and _NAME_. When the _TYPE_ variable has one of the following values, the _NAME_ variable identifies the row of the matrix:

TYPE	Contents
CSSCP	corrected SSCP matrix for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PSSCP	pooled within-class corrected SSCP matrix
BSSCP	between-class SSCP matrix
COV	covariance matrix for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PCOV	pooled within-class covariance matrix
BCOV	between-class covariance matrix
CORR	correlation matrix for the total sample (CLASS variable missing) and within each class (CLASS variable present)
PCORR	pooled within-class correlation matrix
BCORR	between-class correlation matrix

When the _TYPE_ variable has one of the following values, the _NAME_ variable identifies the canonical variable:

TYPE	Contents
CANCORR	canonical correlations
STRUCTUR	canonical structure
BSTRUCT	between canonical structure
PSTRUCT	pooled within-class canonical structure
SCORE	total sample standardized canonical coefficients
PSCORE	pooled within-class standardized canonical coefficients
RAWSCORE	raw canonical coefficients
CANMEAN	means of the canonical variables for each class

You can use this data set in PROC SCORE to get scores on the canonical variables for new data by using one of the following forms:

```
* The CLASS variable C is numeric;
proc score data=NewData score=Coef(where=(c = .  )) out=Scores;
run;

* The CLASS variable C is character;
proc score data=NewData score=Coef(where=(c = ' ')) out=Scores;
run;
```

The WHERE clause is used to exclude the within-class means and standard deviations. PROC SCORE standardizes the new data by subtracting the original variable means that are stored in the _TYPE_='MEAN' observations and dividing by the original variable standard deviations from the _TYPE_='STD' observations. Then PROC SCORE multiplies the standardized variables by the coefficients from the _TYPE_='SCORE' observations to get the canonical scores.

Computational Resources

In the following discussion, let

$$\begin{aligned} n &= \text{number of observations} \\ c &= \text{number of class levels} \\ v &= \text{number of variables in the VAR list} \\ l &= \text{length of the CLASS variable} \end{aligned}$$

Memory Requirements

The amount of memory in bytes for temporary storage needed to process the data is

$$c(4v^2 + 28v + 4l + 68) + 16v^2 + 96v + 4l$$

For the ANOVA option, the temporary storage must be increased by $16v$ bytes. The DISTANCE option requires an additional temporary storage of $4v^2 + 4v$ bytes.

Time Requirements

The following factors determine the time requirements of the CANDISC procedure:

- The time needed for reading the data and computing covariance matrices is proportional to nv^2. PROC CANDISC must also look up each class level in the list. This is faster if the data are sorted by the CLASS variable. The time for looking up class levels is proportional to a value that ranges from n to $n \log(c)$.
- The time for inverting a covariance matrix is proportional to v^3.
- The time required for the canonical discriminant analysis is proportional to v^3.

Each of the preceding factors has a different constant of proportionality.

Displayed Output

The displayed output from PROC CANDISC includes the class level information table. For each level of the classification variable, the following information is provided: the output data set variable name, frequency sum, weight sum, and the proportion of the total sample.

The optional output from PROC CANDISC includes the following:

- Within-class SSCP matrices for each group
- Pooled within-class SSCP matrix
- Between-class SSCP matrix
- Total-sample SSCP matrix
- Within-class covariance matrices for each group
- Pooled within-class covariance matrix
- Between-class covariance matrix, equal to the between-class SSCP matrix divided by $n(c-1)/c$, where n is the number of observations and c is the number of classes
- Total-sample covariance matrix
- Within-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the within-class population correlation coefficients are zero
- Pooled within-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the partial population correlation coefficients are zero
- Between-class correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the between-class population correlation coefficients are zero
- Total-sample correlation coefficients and $\mathrm{Pr} > |r|$ to test the hypothesis that the total population correlation coefficients are zero
- Simple statistics, including N (the number of observations), sum, mean, variance, and standard deviation both for the total sample and within each class
- Total-sample standardized class means, obtained by subtracting the grand mean from each class mean and dividing by the total sample standard deviation
- Pooled within-class standardized class means, obtained by subtracting the grand mean from each class mean and dividing by the pooled within-class standard deviation
- Pairwise squared distances between groups
- Univariate test statistics, including total-sample standard deviations, pooled within-class standard deviations, between-class standard deviations, R square, $R^2/(1-R^2)$, F, and $\mathrm{Pr} > F$ (univariate F values and probability levels for one-way analyses of variance)

By default, PROC CANDISC displays these statistics:

- Multivariate statistics and F approximations, including Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root with F approximations, numerator and denominator degrees of freedom (Num DF and Den DF), and probability values $(\mathrm{Pr} > F)$. Each of these four multivariate statistics tests the hypothesis that the class means are equal in the population. For more information, see the section "Multivariate Tests" on page 98 in Chapter 4, "Introduction to Regression Procedures."

- Canonical correlations

- Adjusted canonical correlations (Lawley 1959). These are asymptotically less biased than the raw correlations and can be negative. The adjusted canonical correlations might not be computable and are displayed as missing values if two canonical correlations are nearly equal or if some are close to zero. A missing value is also displayed if an adjusted canonical correlation is larger than a previous adjusted canonical correlation.

- Approximate standard error of the canonical correlations

- Squared canonical correlations

- Eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$. Each eigenvalue is equal to $\rho^2/(1-\rho^2)$, where ρ^2 is the corresponding squared canonical correlation and can be interpreted as the ratio of between-class variation to pooled within-class variation for the corresponding canonical variable. The table includes eigenvalues, differences between successive eigenvalues, the proportion of the sum of the eigenvalues, and the cumulative proportion.

- Likelihood ratio for the hypothesis that the current canonical correlation and all smaller ones are zero in the population. The likelihood ratio for the hypothesis that all canonical correlations equal zero is Wilks' lambda.

- Approximate F statistic based on Rao's approximation to the distribution of the likelihood ratio (Rao 1973, p. 556; Kshirsagar 1972, p. 326)

- Numerator degrees of freedom (Num DF), denominator degrees of freedom (Den DF), and $\mathrm{Pr} > F$, the probability level associated with the F statistic

You can suppress the following statistics by specifying the SHORT option:

- Total canonical structure, giving total-sample correlations between the canonical variables and the original variables

- Between canonical structure, giving between-class correlations between the canonical variables and the original variables

- Pooled within canonical structure, giving pooled within-class correlations between the canonical variables and the original variables

- Total-sample standardized canonical coefficients, standardized to give canonical variables that have zero mean and unit pooled within-class variance when applied to the total-sample standardized variables

- Pooled within-class standardized canonical coefficients, standardized to give canonical variables that have zero mean and unit pooled within-class variance when applied to the pooled within-class standardized variables

- Raw canonical coefficients, standardized to give canonical variables that have zero mean and unit pooled within-class variance when applied to the centered variables
- Class means on the canonical variables

ODS Table Names

PROC CANDISC assigns a name to each table that it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 35.2. For more information about ODS, see Chapter 23, "Using the Output Delivery System."

Table 35.2 ODS Tables Produced by PROC CANDISC

ODS Table Name	Description	PROC CANDISC Option
ANOVA	Univariate statistics	ANOVA
AveRSquare	Average R square	ANOVA
BCorr	Between-class correlations	BCORR
BCov	Between-class covariances	BCOV
BSSCP	Between-class SSCP matrix	BSSCP
BStruc	Between canonical structure	Default
CanCorr	Canonical correlations	Default
CanonicalMeans	Class means on canonical variables	Default
Counts	Number of observations, variables, classes, degrees of freedom	Default
CovDF	Degrees of freedom for covariance matrices, not printed	Any *COV option
Dist	Squared distances	DISTANCE
DistFValues	*F* statistics based on squared distances	DISTANCE
DistProb	Probabilities for *F* statistics from squared distances	DISTANCE
Levels	Class level information	Default
MultStat	MANOVA	Default
NObs	Number of observations	Default
PCoef	Pooled standard canonical coefficients	Default
PCorr	Pooled within-class correlations	PCORR
PCov	Pooled within-class covariances	PCOV
PSSCP	Pooled within-class SSCP matrix	PSSCP
PStdMeans	Pooled standardized class means	STDMEAN
PStruc	Pooled within canonical structure	Default
RCoef	Raw canonical coefficients	Default
SimpleStatistics	Simple statistics	SIMPLE
TCoef	Total-sample standard canonical coefficients	Default
TCorr	Total-sample correlations	TCORR
TCov	Total-sample covariances	TCOV

Table 35.2 *continued*

ODS Table Name	Description	PROC CANDISC Option
TSSCP	Total-sample SSCP matrix	TSSCP
TStdMeans	Total standardized class means	STDMEAN
TStruc	Total canonical structure	Default
WCorr	Within-class correlations	WCORR
WCov	Within-class covariances	WCOV
WSSCP	Within-class SSCP matrices	WSSCP

Example: CANDISC Procedure

Example 35.1: Analyzing Iris Data by Using PROC CANDISC

The iris data that were published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in millimeters in 50 iris specimens from each of three species: *Iris setosa*, *I. versicolor*, and *I. virginica*. The iris data set is available from the Sashelp library.

This example is a canonical discriminant analysis that creates an output data set that contains scores on the canonical variables and plots the canonical variables.

The following statements produce Output 35.1.1 through Output 35.1.6:

```
title 'Fisher (1936) Iris Data';

proc candisc data=sashelp.iris out=outcan distance anova;
   class Species;
   var SepalLength SepalWidth PetalLength PetalWidth;
run;
```

PROC CANDISC first displays information about the observations and the classes in the data set in Output 35.1.1.

Output 35.1.1 Iris Data: Summary Information

Fisher (1936) Iris Data

The CANDISC Procedure

Total Sample Size	150	**DF Total**	149
Variables	4	**DF Within Classes**	147
Classes	3	**DF Between Classes**	2

Number of Observations Read	150
Number of Observations Used	150

Output 35.1.1 *continued*

Class Level Information				
Species	**Variable Name**	**Frequency**	**Weight**	**Proportion**
Setosa	Setosa	50	50.0000	0.333333
Versicolor	Versicolor	50	50.0000	0.333333
Virginica	Virginica	50	50.0000	0.333333

The DISTANCE option in the PROC CANDISC statement displays squared Mahalanobis distances between class means. Results from the DISTANCE option are shown in Output 35.1.2.

Output 35.1.2 Iris Data: Squared Mahalanobis Distances and Distance Statistics

Fisher (1936) Iris Data

The CANDISC Procedure

Squared Distance to Species			
From Species	**Setosa**	**Versicolor**	**Virginica**
Setosa	0	89.86419	179.38471
Versicolor	89.86419	0	17.20107
Virginica	179.38471	17.20107	0

F Statistics, NDF=4, DDF=144 for Squared Distance to Species			
From Species	**Setosa**	**Versicolor**	**Virginica**
Setosa	0	550.18889	1098
Versicolor	550.18889	0	105.31265
Virginica	1098	105.31265	0

Prob > Mahalanobis Distance for Squared Distance to Species			
From Species	**Setosa**	**Versicolor**	**Virginica**
Setosa	1.0000	<.0001	<.0001
Versicolor	<.0001	1.0000	<.0001
Virginica	<.0001	<.0001	1.0000

Output 35.1.3 displays univariate and multivariate statistics. The ANOVA option uses univariate statistics to test the hypothesis that the class means are equal. The resulting R-square values range from 0.4008 for SepalWidth to 0.9414 for PetalLength, and each variable is significant at the 0.0001 level. The multivariate test for differences between the classes (which is displayed by default) is also significant at the 0.0001 level; you would expect this from the highly significant univariate test results.

Output 35.1.3 Iris Data: Univariate and Multivariate Statistics

Fisher (1936) Iris Data

The CANDISC Procedure

Univariate Test Statistics

F Statistics, Num DF=2, Den DF=147

Variable	Label	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
SepalLength	Sepal Length (mm)	8.2807	5.1479	7.9506	0.6187	1.6226	119.26	<.0001
SepalWidth	Sepal Width (mm)	4.3587	3.3969	3.3682	0.4008	0.6688	49.16	<.0001
PetalLength	Petal Length (mm)	17.6530	4.3033	20.9070	0.9414	16.0566	1180.16	<.0001
PetalWidth	Petal Width (mm)	7.6224	2.0465	8.9673	0.9289	13.0613	960.01	<.0001

Average R-Square	
Unweighted	0.7224358
Weighted by Variance	0.8689444

Multivariate Statistics and F Approximations

S=2 M=0.5 N=71

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02343863	199.15	8	288	<.0001
Pillai's Trace	1.19189883	53.47	8	290	<.0001
Hotelling-Lawley Trace	32.47732024	582.20	8	203.4	<.0001
Roy's Greatest Root	32.19192920	1166.96	4	145	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Output 35.1.4 displays canonical correlations and eigenvalues. The R square between Can1 and the CLASS variable, 0.969872, is much larger than the corresponding R square for Can2, 0.222027.

Output 35.1.4 Iris Data: Canonical Correlations and Eigenvalues

Fisher (1936) Iris Data

The CANDISC Procedure

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq): Eigenvalue	Difference	Proportion	Cumulative
1	0.984821	0.984508	0.002468	0.969872	32.1919	31.9065	0.9912	0.9912
2	0.471197	0.461445	0.063734	0.222027	0.2854		0.0088	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.02343863	199.15	8	288	<.0001
2	0.77797337	13.79	3	145	<.0001

Output 35.1.5 displays correlations between canonical and original variables.

Output 35.1.5 Iris Data: Correlations between Canonical and Original Variables

Fisher (1936) Iris Data

The CANDISC Procedure

Total Canonical Structure			
Variable	**Label**	**Can1**	**Can2**
SepalLength	Sepal Length (mm)	0.791888	0.217593
SepalWidth	Sepal Width (mm)	-0.530759	0.757989
PetalLength	Petal Length (mm)	0.984951	0.046037
PetalWidth	Petal Width (mm)	0.972812	0.222902

Between Canonical Structure			
Variable	**Label**	**Can1**	**Can2**
SepalLength	Sepal Length (mm)	0.991468	0.130348
SepalWidth	Sepal Width (mm)	-0.825658	0.564171
PetalLength	Petal Length (mm)	0.999750	0.022358
PetalWidth	Petal Width (mm)	0.994044	0.108977

Pooled Within Canonical Structure			
Variable	**Label**	**Can1**	**Can2**
SepalLength	Sepal Length (mm)	0.222596	0.310812
SepalWidth	Sepal Width (mm)	-0.119012	0.863681
PetalLength	Petal Length (mm)	0.706065	0.167701
PetalWidth	Petal Width (mm)	0.633178	0.737242

Output 35.1.6 displays canonical coefficients. The raw canonical coefficients for the first canonical variable, Can1, show that the classes differ most widely on the linear combination of the centered variables: $-0.0829378 \times \text{SepalLength} - 0.153447 \times \text{SepalWidth} + 0.220121 \times \text{PetalLength} + 0.281046 \times \text{PetalWidth}$.

Output 35.1.6 Iris Data: Canonical Coefficients

Fisher (1936) Iris Data

The CANDISC Procedure

Total-Sample Standardized Canonical Coefficients			
Variable	**Label**	**Can1**	**Can2**
SepalLength	Sepal Length (mm)	-0.686779533	0.019958173
SepalWidth	Sepal Width (mm)	-0.668825075	0.943441829
PetalLength	Petal Length (mm)	3.885795047	-1.645118866
PetalWidth	Petal Width (mm)	2.142238715	2.164135931

Output 35.1.6 *continued*

Pooled Within-Class Standardized Canonical Coefficients			
Variable	Label	Can1	Can2
SepalLength	Sepal Length (mm)	-.4269548486	0.0124075316
SepalWidth	Sepal Width (mm)	-.5212416758	0.7352613085
PetalLength	Petal Length (mm)	0.9472572487	-.4010378190
PetalWidth	Petal Width (mm)	0.5751607719	0.5810398645

Raw Canonical Coefficients			
Variable	Label	Can1	Can2
SepalLength	Sepal Length (mm)	-.0829377642	0.0024102149
SepalWidth	Sepal Width (mm)	-.1534473068	0.2164521235
PetalLength	Petal Length (mm)	0.2201211656	-.0931921210
PetalWidth	Petal Width (mm)	0.2810460309	0.2839187853

Output 35.1.7 displays class level means on canonical variables.

Output 35.1.7 Iris Data: Canonical Means

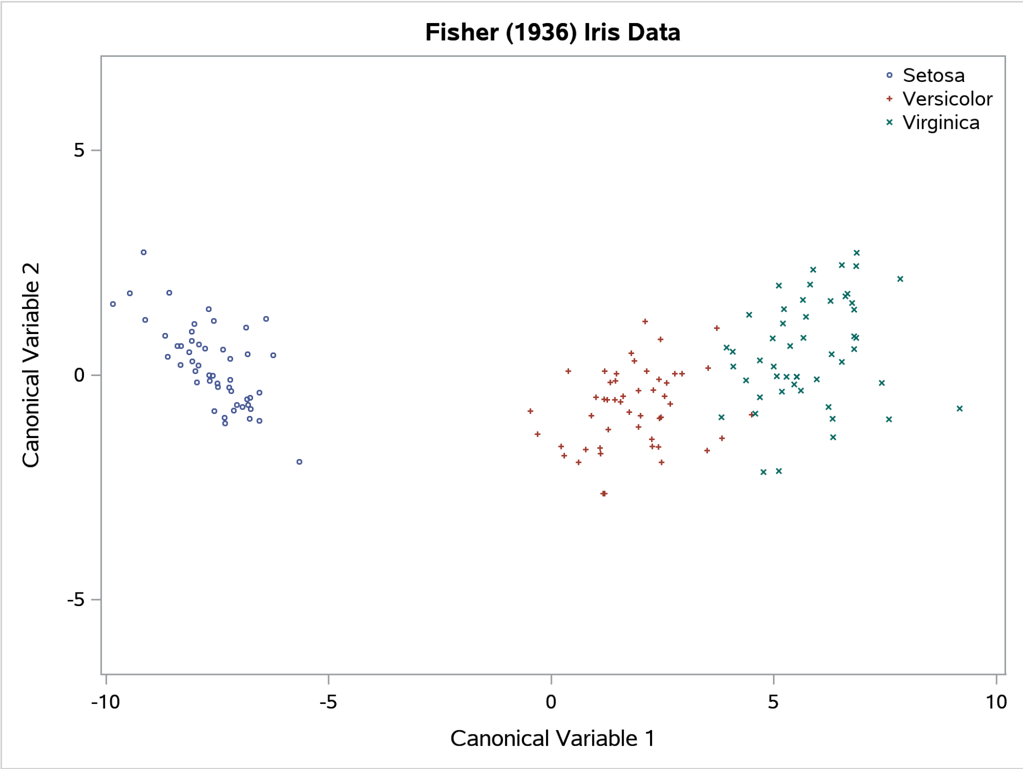
Class Means on Canonical Variables		
Species	Can1	Can2
Setosa	-7.607599927	0.215133017
Versicolor	1.825049490	-0.727899622
Virginica	5.782550437	0.512766605

The TEMPLATE and SGRENDER procedures are used to create a plot of the first two canonical variables. The following statements produce Output 35.1.8:

```
proc template;
   define statgraph scatter;
      begingraph / attrpriority=none;
         entrytitle 'Fisher (1936) Iris Data';
         layout overlayequated / equatetype=fit
            xaxisopts=(label='Canonical Variable 1')
            yaxisopts=(label='Canonical Variable 2');
            scatterplot x=Can1 y=Can2 / group=species name='iris'
                                        markerattrs=(size=3px);
            layout gridded / autoalign=(topright topleft);
               discretelegend 'iris' / border=false opaque=false;
            endlayout;
         endlayout;
      endgraph;
   end;
run;

proc sgrender data=outcan template=scatter;
run;
```

Output 35.1.8 Iris Data: Plot of First Two Canonical Variables



The plot of canonical variables in Output 35.1.8 shows that of the two canonical variables, Can1 has more discriminatory power.

References

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7:179–188.

Kshirsagar, A. M. (1972). *Multivariate Analysis*. New York: Marcel Dekker.

Lawley, D. N. (1959). "Tests of Significance in Canonical Analysis." *Biometrika* 46:59–66.

Puranen, J. (1917). "Fish Catch data set (1917)." Journal of Statistics Education Data Archive. Accessed February 17, 2022. http://jse.amstat.org/jse_data_archive.htm.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. 2nd ed. New York: John Wiley & Sons.